EXPECTED UTILITY FOR QUEUES SERVICING MESSAGES WITH EXPONENTIALLY DECAYING UTILITY

By Frank A. Haight

Institute of Transportation and Traffic Engineering, University of California, Los Angeles

1. Introduction. When a sequence of messages arrives at some center, they may form a queue, owing to delay in reading or processing each message. If the usefulness of a message decays with the lapse of time (as might occur in military operations) it would be important to handle incoming messages with a view to minimizing the loss of utility. In particular, the order in which items are handled assumes a greater importance than in other queueing problems. We consider here a single queue of this sort, and, with some distribution-theoretic restrictions, derive expressions for the expected (terminal) utility in two cases: (a) most recent, and (b) least recent message serviced first, with both random and regular departures.

2. Assumptions. Consider a single queue of messages, in equilibrium, and assume that each message has associated with it at time t after entry, a utility subject to exponential decay. We investigate the loss of utility due to queueing delay in several different circumstances. In each case λ denotes the mean arrival rate, μ the mean departure rate, $\rho = \lambda/\mu$. No messages are removed from the queue without completion of service.

If the initial utility of a message (at the time of entry into the queue) is denoted by y_0, the waiting time in the queue (exclusive of service time) by w and the final utility (when entering service) by y, then we assume y_0 and w to be independent random variables, with

$$y = y_0 e^{-\beta w}, \tag{1}$$

where β is the same for all messages. We also assume the distribution of initial utility to be Type III, i.e.,

$$dF(y_0) = Ke^{-py_0}y_0^{q-1}\,dy_0, \qquad 0 < y_0 < \infty, \tag{2}$$

where $K\Gamma(q) = p^q$. We shall use equations (1) and (2) to determine $E(y)$, and in some circumstances the distribution of y also. If the Laplace transform of the distribution of w is $\phi(s)$, then $E(y) = (q/p)\phi(\beta)$.

3. First come, first served; Poisson service. This means that messages are taken off the bottom of the pile, and that both arrivals and departures occur at random instants. Then the distribution of queue length N (including the message being serviced) is

$$p_n = \text{Prob}\ (N = n) = (1 - \rho)\rho^n. \tag{3}$$

Received October 8, 1959; revised August 22, 1960.

The waiting time of a message will consist of n negative exponential phases with probability p_n. Its distribution therefore consists of a discrete magnitude $p_0 = 1 - \rho$ at the origin, together with the continuous component

$$h(w) = \lambda(1 - \rho)e^{(\lambda-\mu)w}, \qquad 0 < w < \infty. \tag{4}$$

Since y_0 and w are independent, we can now write down the distribution of y. It will consist of the sum of two terms, corresponding to the two components of the distribtion of w. The first of these corresponds to $y = y_0$, occurs with probability $1 - \rho$ and is therefore

$$(1 - \rho)Ke^{-py}y^{q-1}. \tag{5}$$

The second term of the distribution of y is found by integrating y_0 out of the joint distribution of y_0 and y:

$$Ke^{-py_0}y_0^{q-1}\lambda(1 - \rho) \exp [-(\lambda/\mu) \log (y/y_0)](1/\beta y)$$

since $dw = dy/\beta y$. Letting

$$\Gamma(n, x) = \int_x^\infty e^{-t}t^{n-1}\,dt$$

denote the incomplete gamma function, we have for the second term of the distribution of y

$$\frac{\lambda(1 - \rho)}{\beta\Gamma(q)y}(py)^{(\mu-\lambda)/\beta}\,\Gamma\left(q - \frac{\mu}{\beta} + \frac{\lambda}{\beta}, py\right). \tag{6}$$

$E(y)$ can be found easily from this expression, or from the Laplace transform of (4) (with the discrete element added), and turns out to be

$$E(y) = (q/p)\left[1 - \rho + \frac{\lambda(1 - \rho)}{\mu - \lambda + \beta}\right]. \tag{7}$$

If we choose time units so that $\lambda = 1$, and units of utility so that $E(y_0) = (q/p) = 1$, and let $\alpha = (\log 2)/\beta$ denote the half-life of information, then we can tabulate $E(y)$ as a function of service rate and half-life. The values in Table 1 have been obtained.

TABLE 1

Expected terminal utility as percentage of expected initial utility; first come, first served; negative exponential service

	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$	$\alpha = 5$
$\mu = 1$	0	0	0	0	0
$\mu = 2$	.79531	.87131	.91421	.92615	.93912
$\mu = 3$	.91421	.95077	.96548	.97342	.97792
$\mu = 4$	.95308	.97411	.98212	.98635	.98896
$\mu = 5$	.97046	.98405	.98908	.99170	.99330

TABLE 2

Expected terminal utility as percentage of expected initial utility; first come, first served; regular service

	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$	$\alpha = 5$
$\mu = 1$	0	0	0	0	0
$\mu = 2$	.86591	.92437	.94715	.95921	.96631
$\mu = 3$	.94924	.97286	.98148	.98632	.98866
$\mu = 4$	.97345	.98602	.99084	.99281	.99466
$\mu = 5$	.98366	.99136	.99430	.99599	.99635

4. First come, first served; regular departures. In case service is regular rather than Poisson, we need only replace (3) by a formula quoted by Saaty [2], p. 177, formula (16). The calculations leading to the distribution of y are then rather cumbersome, and will be omitted. We can obtain the mean value of this quantity by the simpler method, using the formula for the Laplace transform of $h(w)$ quoted by Kendall [5], p. 156, formula (16):

$$\phi(s) = \frac{(s/\mu)(1 - \rho)}{(s/\mu) + \rho(e^{-s/\mu} - 1)}\,. \tag{8}$$

Making the same choice of units as in the previous section, the values in Table 2 are obtained.

5. Last come, first served; regular departures. In this case messages are taken off the top of the pile. The number of service phases (each of length $1/\mu$) which will delay a given message is by no means the number of waiting messages encountered. However, the probability of an empty queue is the same, and therefore the first term of the distribution of y is given by (5) in this case also.

To find the second term of the distribution, consider a new arrival to the queue, which is not empty. Define an auxiliary queue to consist of this arrival and all subsequent arrivals. Then the probability π_n, $n = 1, 2, \cdots$ that the original arrival will be preceded into service by n other messages (including the one being serviced at his arrival time) is the probability that the auxiliary (beginning with one member) will discharge exactly n messages before first becoming empty.

Therefore the continuous component of the distribution of w is

$$h(w) = \mu\rho\pi_n, \qquad (n-1)/\mu < w < n/\mu, \qquad n = 1, 2, \cdots, \tag{9}$$

and of y for fixed y_0 is

$$h(y \mid y_0) = \frac{\lambda\pi_n}{\beta y}, \qquad y_0 \exp\,[\beta\,(n-1)/\mu] > y > y_0 \exp\,(-n\beta/\mu), \qquad n = 1, 2, \cdots. \tag{10}$$

An extra factor ρ has been introduced into (9) which was not required in the corresponding formula of Section 3. This is because the distribution π_n is defined over the integers excluding zero, whereas the queue length distribution (3) is defined over all non-negative integers. Thus the required integral

$$\int_0^\infty h(w)\,dw = \rho$$

obtains automatically in that case, but must be produced artificially in this one.

Denoting the second term of the density of y by $f_2(y)$, its contribution to the expected value of y by $E_2(y)$, and making the convenient abbreviation $a_n = \exp(n\beta/\mu)$, we have

$$(11)\qquad f_2(y) = \sum_{n=1}^{\infty} \int_{a_{n-1}y}^{a_n y} Ke^{-py_0} y_0^{q-1} \frac{\lambda \pi_n}{\beta y}\,dy_0,$$

and therefore

$$(12)\qquad E_2(y) = \int_0^\infty \sum_{n=1}^{\infty} \int_{a_{n-1}y}^{a_n y} Ke^{-py_0} y_0^{q-1} \frac{\lambda \pi_n}{\beta}\,dy_0\,dy.$$

Using the transformation

$$z = (y_0 - a_{n-1}y)/(y_0 - a_n y)$$

from y_0 to z, we obtain after integration

$$(13)\qquad E_2(y) = (\lambda q/\beta p)(e^{\beta/\mu} - 1)\pi(e^{-\beta/\mu}),$$

where $\pi(s) = \sum \pi_n s^n$ is the probability generating function of the π_n distribution.

Now we consider the π_n distribution itself, and its generating function $\pi(s)$. The most unfavorable situation for an entry into the queue is that taking place just at the beginning of a service time. For this case Borel [1] has given the value for π_n, namely

$$\pi_n = [n^{n-2}/\Gamma(n)]e^{-\rho n}\rho^{n-1},$$

for which the generating function is (cf., Haight and Breuer [3])

$$\pi(s) = s \exp \sum_{n=1}^{\infty} \frac{n^{n-1}\rho^n}{n!\,e^{\rho n}}(s^n - 1).$$

TABLE 3

Approximate expected terminal utility as percentage of expected initial utility; last come, first served; regular service

	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$	$\alpha = 5$
$\mu = 1$	.33466	.45492	.52272	.56822	.59057
$\mu = 2$	.84089	.90260	.92936	.94446	.95434
$\mu = 3$	.93811	.96572	.97626	.98206	.98530
$\mu = 4$	.96835	.98304	.98866	.99115	.99331
$\mu = 5$	.98098	.99006	.99326	.99520	.99574

Using this approximation, we find that

$$E_2(y) \approx (\lambda q/\beta p)(1 - e^{-\beta/\mu}) \exp(\sigma - \rho), \tag{14}$$

where $\sigma e^{-\sigma} = \exp[-\rho - (\beta/\mu)]$. With the same choice of units and definition of α, formula (14) yields the values in Table 3. Next, we obtain exact values for π_n; this constitutes one generalization of Borel's distribution. Another generalization is that of Tanner [7], [3] and still another will appear in the following section.

Let t be the time remaining between an arrival and the first service termination; this quantity is rectangularly distributed over the interval $(0, 1/\mu)$. The probability of x arrivals in the time t is therefore

$$\int_0^{1/\mu} \mu \frac{(\lambda t)^x}{x!} e^{-\lambda t}\, dt = \frac{1}{\rho} \frac{\gamma(x + 1, \rho)}{\Gamma(x + 1)}, \tag{15}$$

where $\Gamma(n) = \Gamma(n, z) + \gamma(n, z)$. If, in addition to t, any complete service periods must be waited, the probabilities of x arrivals during one of these are the simple Poisson expressions. In order that the queue beginning with one member shall vanish for the first time when exactly n members have passed through it requires exactly $n - 1$ arrivals in the n service periods (including the fractional one), subject to the restrictions that there will be no arrivals in the last period, no more than one in the next to last, and so forth. As an occupancy problem, we want to put $n - 1$ balls into $n - 1$ boxes so that, reading from left to right at least as many balls are passed as boxes.

Combining (15) for the fractional period with the Poisson terms for the whole periods in this way we obtain

$$\pi_1 = \gamma(1, \rho)/\rho,$$

$$\pi_2 = \gamma(2, \rho)/\rho e^{\rho},$$

$$\pi_3 = [\gamma(3, \rho) + 2\rho\gamma(2, \rho)]/2\rho e^{2\rho},$$

and in general

$$\pi_n = \frac{\sum_{i=2}^{n} \binom{n-2}{i-2} \rho^{n-i}\gamma(i, \rho)(n - 1)^{n-i}}{\Gamma(n)\rho e^{(n-1)\rho}}. \tag{16}$$

To find the generating function for this distribution, multiply π_n by s^n and add the terms containing $\gamma(i, \rho)$ separately for each i. The first of these is simply $(s/\rho)\gamma(1, \rho)$. The succeeding values of i yield infinite series of the type mentioned by Bromwich [6], p. 160, example 4. If we write the nth sum in the form

$$\Delta_n[s^n\gamma(n, \rho)]/[\Gamma(n)e^{(n-1)}] \tag{17}$$

then, using the example of Bromwich, the Δ_n satisfy the equation

$$(s\rho e^{-\rho})^{n-1}\Delta_n = [(\log \Delta_n)/(n - 1)]^{n-1}. \tag{18}$$

Hence the generating function can be written

$$\pi(s) = \frac{e^{\rho}}{\rho}\sum_{n=1}^{\infty}\frac{\gamma(n, \rho)}{\Gamma(n)}(se^{-\rho})^{n}\Delta_{n}. \tag{19}$$

An attempt to use (19) with (13) to compute $E_2(y)$ leads to very substantial calculations, most particularly in connection with finding the Δ_n from (18).

The short method of Laplace transforms also gives an expression for $E_2(y)$ which is awkward to compute. Tanner [7] gives the Laplace transform of the delay in the form $\phi(s) = \sigma^{\mu t}$, where t is the fractional service period and σ satisfies

$$\log\sigma = \rho(\sigma - 1) + (s/\mu). \tag{20}$$

Averaging over t, we find

$$E_2(y) = \frac{\rho(\sigma - 1)}{\rho(\sigma - 1) - (\beta/\mu)}, \tag{21}$$

where

$$\log\sigma = \rho(\sigma - 1) + (\beta/\mu). \tag{22}$$

6. Last come, first served; Poisson service. To deal with this case we need first a formula generalizing Borel's distribution to negative exponential service time. Since departures are random, there is now no distinction to be observed regarding the fractional service time on entry to the queue. The probability of x arrivals in a single service interval is

$$\int_0^{\infty}\mu e^{-\mu v}(1/x!)(\lambda v)^{x}e^{-\lambda v}\,dv = \rho^{x}(1 + \rho)^{-x-1}. \tag{23}$$

Thus π_n are all of the form

$$\pi_n = K_n[\rho^{n-1}/(1 + \rho)^{2n-1}], \qquad n = 1, 2, \cdots, \tag{24}$$

where K_n represents the number of ways these arrivals can occur subject to the restrictions mentioned in the last section.

The values of K_n can be found by use of Cauchy's theorem in much the same way as Borel used the theorem to evaluate the coefficients in the simpler case Bateman [4] also refers to these numbers (p. 230) in a different context; both methods yield the expression

$$K_n = \frac{1}{n}\binom{2n-2}{n-1}. \tag{25}$$

Bateman also gives the generating function of the K_n

$$G(s) = \sum_{n=1}^{\infty}K_n s^{n} = \tfrac{1}{2} - \tfrac{1}{2}(1 - 4s)^{\frac{1}{2}}, \tag{26}$$

which is useful in finding $E_2(y)$.

Given n, there must have been n departures and $n - 1$ arrivals between the arrival of the particular message and its entry into service. The spacings between these $2n - 1$ events are each distributed with density $(\lambda + \mu) \exp -(\lambda + \mu)x$, and therefore

$$h(w) = \sum_{n=1}^{\infty} \frac{1}{n} \binom{2n-2}{n-1} \frac{\rho^n}{(1+\rho)^{2n-1}} \frac{(\lambda+\mu)^{2n-1} w^{2n-2}}{\Gamma(2n-1) e^{(\lambda+\mu)w}} \tag{27}$$
$$= e^{-(\lambda+\mu)w} \sum_{n=1}^{\infty} \frac{1}{n} \binom{2n-2}{n-1} \frac{\lambda^n \mu^{n-1} w^{2n-2}}{(2n-2)!},$$

leading to

$$\phi(\beta) = \int_0^{\infty} e^{-\beta w} e^{-(\lambda+\mu)w} \sum_{n=1}^{\infty} \frac{1}{n} \binom{2n-2}{n-1} \frac{\lambda^n \mu^{n-1} w^{2n-2}}{(2n-2)!} \, dw$$

or, in terms of the generating function $G(s)$,

$$\phi(\beta) = [(\lambda + \mu + \beta)/\mu] G[\lambda\mu/(\lambda + \mu + \beta)^2]. \tag{28}$$

Using (26), we obtain

$$\phi(\beta) = E_2(y) = [(\lambda + \mu + \beta)/\mu]\{\tfrac{1}{2} - \tfrac{1}{2}\{1 - 4\lambda\mu/(\lambda + \mu + \beta)^2\}^{\frac{1}{2}}\}. \tag{29}$$

Formula (29) has been used in computing Table 4.

TABLE 4

Expected terminal utility as percentage of expected initial utility; last come, first served; negative exponential service

	$\alpha = 1$	$\alpha = 2$	$\alpha = 3$	$\alpha = 4$	$\alpha = 5$
$\mu = 1$	.44476	.55964	.62110	.65929	.69059
$\mu = 2$	.82961	.88946	.91727	.93366	.94453
$\mu = 3$	.92114	.95353	.96695	.97433	.97902
$\mu = 4$	.95524	.97486	.98232	.98657	.98911
$\mu = 5$	.97133	.98434	.98922	.99177	.99335

I am grateful to the referee for pointing out a number of errors in earlier versions of this paper, and to Mr. John Riordan for confirming formula (16).

REFERENCES

[1] Borel, Émile, "Sur l'emploi du théorème de Bernoulli pour faciliter le calcul d'une infinité de coefficients. Application au problème de l'attente à un guichet," *Comtes Rendus Acad. Sci. (Paris)*, Vol. 214 (1942), pp. 452–6.

[2] Saaty, Thomas L., "Résumé of useful formulas in queueing theory," *Operations Research*, Vol. 5 (1957), pp. 161–200.

[3] Haight, F. A. and Breuer, M. A., "The Borel-Tanner distribution," *Biometrika*, Parts 1 and 2, Vol. 47 (1960), pp. 143–150.

[4] Erdelyi, *et al*, (Bateman Manuscript Project) *Higher Transcendental Functions*, Vol. III. McGraw Hill Book Co., New York, 1955.

[5] Kendall, D. G., "Some problems in the theory of queues," *J. Roy. Stat. Soc.*, Ser. B, Vol. 13 (1951), pp. 151–185.

[6] Bromwich, T. J. I'A., *An Introduction to the Theory of Infinite Series*, Macmillan and Co., Ltd., London, 1955.

[7] Tanner, J. C., "A problem of interference between two queues." *Biometrika*, Vol. 40 (1953), pp. 58–69.